2G1-01

判例文の論理展開と意味的類似性に基く法律文要約手法の検討

Extracting important sentences from legal precedents based on sematic correlations of words and syntactic phrases.

大島　敦史[*1]　　原口　誠[*1]
Atsushi Ohshima　Makoto Haraguchi

[*1]北海道大学大学院情報科学研究科
Graduate School of Infor.Sci.&Tech., Hokkaido University

We proposes an automatic summarization system that extracts important sentences from a given legal precedent sentence. Those sentences describe legal facts, to be conrresponded to the requirements of legal norm, and relate those ones to some judgements or decisions. So, we propse here a PageRank model to evaluate each sentence from the semantic correspondence between the legal facts and the requirements and from some judgemental sentences involving syntactic key phrases about the decisions.

1　はじめに

近年、コンピューターによって処理された膨大な量の情報が存在している。その膨大な量の情報としてテキスト情報が上げられる。インターネットの発達、様々な文書の電子化によってテキスト情報はどんどん増えていく。しかし、その膨大な量のテキスト情報からどのようにして有用な情報を取り出すかが非常に重要な問題として挙げられる。

法律という分野においても、法律文は電子化されてきておりコンピューター上での様々な処理が容易となっている。代表的な法律文として判例文も裁判所、企業等で電子化されたデータベースによって必要な文献が容易に検索できるようになっている。

しかし、判例文データが膨大な数になると適切な検索条件を選ばなければ、また適切な検索条件を知っていなければ、ユーザーに対して膨大な数の検索結果を検索システムは出力してしまう。

そこで一般的な判例文検索システムでは、必要な判例文をユーザーが直感的に判断できるように判例の要旨を同時に出力している。要旨は法律分野という専門性の高さから法律問題に詳しい専門家が作成することになる。しかし、法理の定説、学説、他の判例文等を勘案し、要約対象とする判例における法的に重要な部分を探さなければいけないため時間的コスト、金銭コストがかかってしまう。

そこで、本研究では判例文から文を単位とする抜粋による自動要旨作成システムを構築する。

2　要旨に必要な条件

本システムで求めたい要旨に必要な条件は以下になる。

1. 原文よりも少ない文量であること。
2. 原文の情報を可能な限り残していること。
3. 「要件事実」を含んでいること。

連絡先: 大島敦史 (4/1 より (株) ＮＴＴドコモ北海道所属), 北海道大学情報科学研究科知識ベース研究室, 060-0814 札幌市北区北 14 西 9, 011-706-7161, {oshima, makoto}@kb.ist.hokudai.ac.jp

本研究ではこの「要件事実」を抽出できる要約システムを提案する。

21　「要件事実」とは

特定の法条文を参照して判例文が書かれているということを上で述べたが、判例文中に書かれている条文の具体的事実を「要件事実」と呼んでいる [加賀山 99]。例えば、「虚偽な点があれば契約は無効」というような法条文が参照されているとき、「虚偽な点」を具体的に示す事実が「要件事実」ということになる。

22　「要件事実」の特徴

本システムに取り入れるべき「要件事実」の特徴を述べる。「要件事実」は抽象度が高い表現で記述されている条文の具体例である。そこで、条文中に使われている語と「要件事実」文中に使われている語の間には強い関連性があると考えられる。例えば、「虚偽な点があれば契約は無効」という条文に対して「A さんは B さんに仮装の取引を持ちかけた」という「要件事実」が対応しているとすると、このとき「虚偽」と「仮装」という語は非常に似ており、「仮装」という語が「虚偽」であることを直接示しているので、「仮装」という単語はこの条文に対応する「要件事実」を特徴付ける単語ということができる。

さらに、「A さんの偽装の工作は C さんの指示によるものだ」という話が同時に記述されているとすると、上の場合と同じように「虚偽」と「偽装」という語にも類義性が見うけられるが、「仮装」と「偽装」という語間にも同じように類義性が見てとれ、「偽装」も特徴的な語ということができる。そこで、「A さんは B さんに仮装の取引を持ちかけた」という話と「A さんの偽装の工作は C さんの指示によるものだ」という話の間にストーリー性があると考えると、それは条文に使われている「虚偽」という語と語義が非常に近い「仮装」と「偽装」というある条文に対して特徴的な語を介した語に文間の連結性によるものであると考える。

また別の特徴として、判例文の語尾表現に着目する。「〜が認められる。」などの語尾表現が用いられている文は事件に対する裁判所の判断を下している文である場合が多い。事件の当事者が主張する「要件事実」が条文に適用している（または、適用していない）といった判断を裁判所は下しているよって、裁判所が何らかの判断を下している文は必要な「要件事実」である可能性が高いと考えられる。

3 システムの設計

「要件事実」を抽出するために、

- ある条文で話題となる語で「要件事実」を特徴付ける。
- 特徴的な語を介した連結性により「要件事実」のストーリー性を捉える。
- 語尾表現を手がかり語として「要件事実」を抽出する。

この 3 点がシステムに要求される。

条文に使われている語と、「要件事実」に使われている語の関連性については述べた。そこで、特定の条文が話題となっている文書群では条文で使われている語と同時に、関連性の高い語も頻出することが予想される。そこで、条文が話題となっている文書群としてある条文名で検索した web ページを用いて、コーパスとし特徴語を抽出した。

文のストーリー性を捉える要約アルゴリズムとして web ページをランキング化する pagerank を応用した方法が研究されている [四ツ谷 03]。この研究では文をノード、文間の関係を共起語の重みを反映させた重みつきリンクと考えて、

$$R_{i+1} = MR_i \tag{1}$$

という再帰的な文の重み付けをすることで、重要度 R を求める。ただし、M は文間の状態推移行列であるノードから別のあるノードに推移する確率は別のノードに移行する量（リンク重み）をあるノードの全流入量で割った値となる。web で利用される pagerank には「定期的にまったく関係のない状態にジャンプする」というブラウジングモデルを数式化して

$$R_{i+1} = (1-\alpha)\, MR_i + \alpha \left[\frac{1}{N}\right]_{N\times 1} \tag{2}$$

ただし、α は経験的に 0.1 〜 0.2 の範囲であり、N はページ数をあらわしている。語尾表現が裁判所の判断を表していると思われる文は「要件事実」文である可能性が高い。そこでこのブラウジングモデルを利用して「重要表現がある文には定期的にユーザーの目が向く」というモデルを考える。

$$R_{i+1} = (1-\alpha)\, MR_i + \alpha\vec{p} \tag{3}$$

ただし、$\vec{p}$ は文の数だけ存在する行ベクトルとし、語尾表現の重要だと思われる文をあらわす行にのみ $\frac{1}{N_{\text{語尾}}}$（ただし、$N_{\text{語尾}}$ は語尾表現が重要であると思われる文数。）で表現される。重要な語尾に関しては今のとこと人手で選定し、「〜が認められる。」といったように「認」という表現を用いているものや、「〜と判断する。」といったように「判断」といった表現を用いているものを考えた。

本研究では「要件事実」のストーリを捉えるために pagerank アルゴリズムを採用する。しかし、文間のリンク重みは特徴語で文をベクトル表現し、文間の以下の余弦類似尺度とする。

$$W_{S_1\to S_2} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4}$$

ここで、$W_{S_1\to S_2}$ はリンクの強度を表し、x_i と y_i は文 S_1、S_2 に含まれる語の構成されている単語番号 i 番目の単語である。重要な「要件事実」文同士は非常に似た、もしくは意味的に似ている語で構成されていることによる。

「要件事実」文同士は意味的な類義語で文間の意味的なストーリー性があると推察する。抽出した特徴語同士の意味的類似性を捉えるために、潜在的意味解析を行う。潜在的意味解析は類義語を同一の次元に縮約することで、類義語の発見を可能にしている [大内 03]。そこで、特徴語の抽出時に用いた web コーパスを文と単語のベクトルで表し特異値分解を行う。データ行列 D は web コーパスの文単語ベクトルである。$m \times n$ 行列の特異値分解は以下のようになる。

$$D = U\Sigma V^T \tag{5}$$

ただし、U は $m\times r$ 直交行列（すなわち $U^TU = UU^T = I$）、V は $n\times r$ 直交行列（$V^TV = VV^T = I$）であり、$r = rank(D)$ である。U と V を以下の列ベクトルであらわす。

$$U = [u_1 u_2 \dots u_r] \tag{6}$$

$$V = [u_1 u_2 \dots u_r] \tag{7}$$

判例文中の構成文を特徴語でベクトル表現し、文番号 i の判例文を構成単語の行ベクトル S_i でベクトル表現したとする。文番号 i の文 S_i を web コーパスの特異値分解の結果から k 次元に縮約する例を以下に記す。m 次元の判例文 S_i を web コーパスの k 次元空間に射影したベクトルを $\hat{S}_i$ とすると、

$$\hat{S}_i = \frac{1}{\sigma_i} S_i^T u_i \,(i = 1, 2, \dots, k) \tag{8}$$

で表すことができる。

4 実験結果

東京地方裁判所判決／平成３年（行ウ）第４２号の事例に対して、民法９４条第２項を適用した判決理由文の要約を試みたので、その概要を記す。一般に文長が長すぎ、文抽出の意味での要約は困難であることから、文の分解規則を導入し、２５文に分割し、要約率 20%での実験を行い、簡易要約器として著名な Posum との比較を行った。

ただし、要約の評価は、判例文添付の判示事項との余弦尺度による類似度を用いた。

モデル	抽出文番号	評価値
Posum	7,8,9,12,22	0.26
本方式	7,10,11,13,24	0.41

９４条第２項の要件に対する直接的な要件事実の言明は、7,10,13 であり、本方式はこれらを全て含む故に最低限の目的は達成しているといえる。文を解析する際の指標となる、重要な語尾を持つ文は９文あり、選択された５文は全てそうした文である。逆に言えば、表現上は重要だが主題との関連性が低い４つの文のランクを下げる効果を確認できた。重要でない語尾を持つ文の中には、行為の背景となる重要な理由を示す文もあり、直接的には法律要件とは関係しないが、事件の背景として特筆すべきものも含まれており、要件事実文を確定した後処理で、そうしたものを拾いあげる操作も必要になると思われる。

参考文献

[加賀山 99] 加賀山茂, 要件事実の考え方-大陸法と英米法の考え方の融合をめざして, 名古屋大学大学院法学研究科 (1999).

[四ツ谷 03] 四ツ谷雅輝, 共起語を介した文間の相互依存関係に基づく重要文抽出法の提案, 北海道大学大学院工学研究科, master's thesis (2003).

[大内 03] 大内浩二 and 三浦孝夫 and 塩谷勇, 多義性を考慮した文書検索, DEWS2003, March(2003)